PKU-ICST at TRECVID 2012: Known-item Search Task

Yuxin Peng, Yunbo Peng, Xiaohua Zhai,

Jian Zhang, Tianjun Xiao, Xin Huang, and Kang Cai

Institute of Computer Science and Technology,

Peking University, Beijing 100871, China.

pengyuxin@pku.edu.cn

Abstract

We participate in all two types of known-item search task of TRECVID 2012: automatic search and interactive search. This paper presents our approaches and results. We adopt three kinds of text information, which are XML documents, ASR and OCR. And we index and search the three kinds of pre-processed text individually with Lucene. In addition, the results are combined and re-ranked by two re-ranking approaches. We achieve the good performances, and official evaluation shows that our team is ranked 1st in both automatic search and interactive search.

1 Overview

In known-item search task of TRECVID 2012, we participate in all two types: automatic search and interactive search. We submitted 4 runs, including 3 runs for automatic search, and 1 run for interactive search. The evaluation results of our 4 runs are shown in Table 1. Our team is ranked 1st in both automatic search and interactive search.

Table 1: Results of our submitted 4 runs on KIS task of TRECVID 2012.

Type	ID	Mean Inverted Rank	Brief description
Automatic	F_A_YES_PKU-ICST-MIPL_2	**0.419**	F_A_YES_PKU-ICST-MIPL_3+ Query Expansion
	F_A_YES_PKU-ICST-MIPL_3	0.317	F_A_YES_PKU-ICST-MIPL_4+ Re-ranking
	F_A_YES_PKU-ICST-MIPL_4	0.313	XML Documents+ASR+OCR
Interactive	I_A_YES_PKU-ICST-MIPL_1	**0.792**	F_A_YES_PKU-ICST-MIPL_2

The 4 runs are described as follows:

- F_A_YES_PKU-ICST-MIPL_4: three kinds of text information are adopted, which are XML documents, ASR and OCR.
- F_A_YES_PKU-ICST-MIPL_3: re-rank the results of F_A_YES_PKU-ICST-MIPL_4 by two re-ranking approaches.
- F_A_YES_PKU-ICST-MIPL_2: adding query expansion to F_A_YES_PKU-ICST-MIPL _3.
- I_A_YES_PKU-ICST-MIPL_1: human feedback based on the results returned by F_A_YES_PKU-ICST-MIPL_2, which is the interactive search.

The framework of our KIS system of TRECVID 2012 is shown in Figure 1.

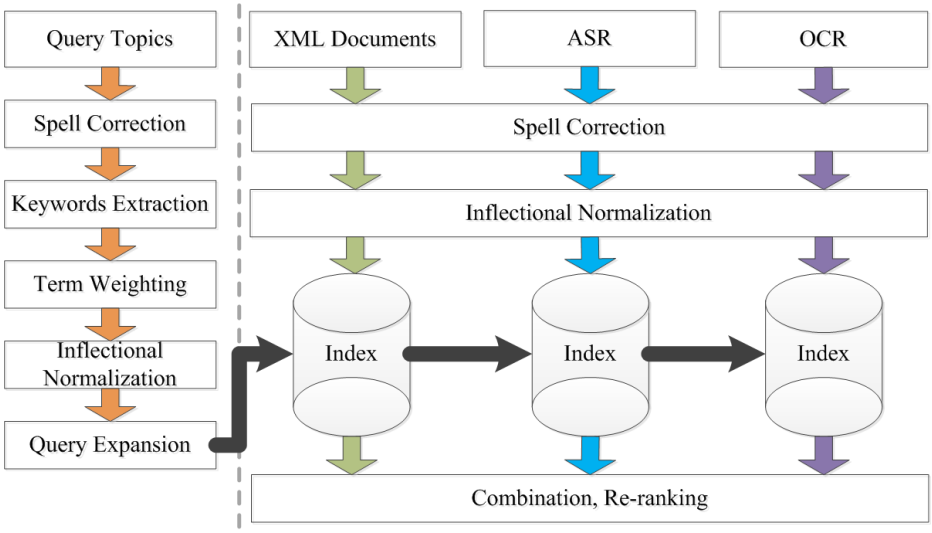


Figure 1: Framework of our KIS system.

2 Text-based Search

We adopt three kinds of text information, which are XML documents, ASR and OCR. In XML documents aspect, we process the XML documents and topics by using the following approaches: spell correction, POS-based keyword extraction, topics term weighting, inflectional normalization and query expansion. These approaches are described in detail as follows.

(1) **Spell Correction:** We find that many words in topics and XML documents are spelt incorrectly, which decrease the performance. We correct the miss-spelt words in topics and XML documents using Aspell [5]. We get the first correction suggestion and append it onto the topics and XML documents [1].

For example, the ground truth of topic 0901 in description region is "*a simple stop motion animation using paperl, mark making tools and saw dust*". The documents will return correctly if "*paper*" is appended according to "*paperl*". Some missing spaces between words are also corrected, for example, topic 1130 "*Find the video showing bursts of white light on mostly black backgrounds which then shows those bursts somewhat clearer showing faint pictures of people and the title "Tracegarden #11" at the end.*". "*Tracegarden*" is split into "*trace garden*".

(2) **POS-based Keywords Extraction:** To extract the keywords, we use Stanford Parser [2] to get

the POS tag and train the weights of different tags by the data of last year. Since the major sentences of XML documents are incomplete, we skip this processing step for XML documents and do this only for topics.

For example, topic 0947 "*Find the video that shows people sitting on a stage in a panel, a large movie screen and a woman wearing a brown and white sweater at a podium with a sign saying Wizards in blue.*" is tagged as below: "*Find/VB the/DT video/NN that/WDT shows/VBZ people/NNS sitting/VBG on/IN a/DT stage/NN in/IN a/DT panel/NN ,/, a/DT large/JJ movie/NN screen/NN and/CC a/DT woman/NN wearing/VBG a/DT brown/JJ and/CC white/JJ sweater/NN at/IN a/DT podium/NN with/IN a/DT sign/NN saying/VBG Wizards/NNP in/IN blue/NN ./.*" We can easily find that the word "*Wizards*" is more important than other words like "*saying*", "*large*", "*panel*" etc.

(3) **Topics Term Weighting:** Since all the KIS topics in TRECVID 2012 are available, so all the topics are regarded as a document collection and the number of topics containing this term are also considered in TF-IDF scheme. The detail of our algorithm is described in Figure 2.

a) We use cosine distance to calculate the relevant score:

$$score(q,d) = \frac{\overrightarrow{V(q)} \cdot \overrightarrow{V(d)}}{\left|\overrightarrow{V(q)}\right| \left|\overrightarrow{V(d)}\right|} \quad (1)$$

where q is topic, d is XML document, $V(q)$ is the weight vector of query q and $V(d)$ is the weight vector of document d. The definition of $\overrightarrow{V(q)}$, $\overrightarrow{V(d)}$, $\left|\overrightarrow{V(q)}\right|$, $\left|\overrightarrow{V(d)}\right|$ are described in step b).

b) We use TF-IDF scheme to compute $V(q)$ and $V(d)$. Different from the default formula given by Lucene, we also consider the number of topics containing this term. Thus:

$$\overrightarrow{V(q)} = \sum_{t\ in\ q} w(t,q) = \sum_{t\ in\ q} tf(t,q) \times idf(t,D) \times idf(t,Q) \quad (2)$$

$$\overrightarrow{V(d)} = \sum_{t\ in\ d} w(t,d) = \sum_{t\ in\ d} tf(t,d) \times idf(t,D) \quad (3)$$

$$\left|\overrightarrow{V(q)}\right| = \sqrt{(\sum_{t\ in\ q} tf(t,q) \times idf(t,D) \times idf(t,Q))^2} \quad (4)$$

$$\left|\overrightarrow{V(d)}\right| = \sqrt{\sum_{t\ in\ d} w(t,d)^2} = \sqrt{\sum_{t\ in\ d} 1^2} = \sqrt{|d|} \quad (5)$$

where D is the set of all XML documents, Q is the set of all topics and we ignore the weight of each term for $\left|\overrightarrow{V(d)}\right|$.

c) Finally, we get the scoring formula:

$$score(q,d) = \frac{\sum_{t\ in\ q}\left(tf(t,d) \times tf(t,q) \times idf(t,D)^2 \times idf(t,Q)\right)}{\sqrt{\left(\sum_{t\ in\ q} tf(t,q) \times idf(t,D) \times idf(t,Q)\right)^2 \times |d|}} \quad (6)$$

Figure 2: Topic term weighting algorithm.

(4) **Inflectional Normalization:** Due to few words exist in XML documents regions (<title>, <description>, <subject>, <keyword>, <keywords>), the information provided by XML documents is limited. To match more words, we get the inflectional normalization of words in topics and XML documents with a well-formed dictionary from [6]. We store the possible transformation of words in topics into a look-up table, and obtain the original form of words

in both topics and XML documents.

(5) **Query Expansion:** The ontology is constructed to expand the query topics.

In ASR aspect, we use the donated ASR data [3]. The processing steps of ASR data are similar to XML documents, including spell correction, POS-based keyword extraction, topics term weighting and inflectional normalization. In OCR aspect, we automatically detect and recognize the text in videos. Then, we do the same processing steps as ASR. Finally, these three kinds of text information are indexed with Lucene individually, and then we parse the topics and get the retrieval results.

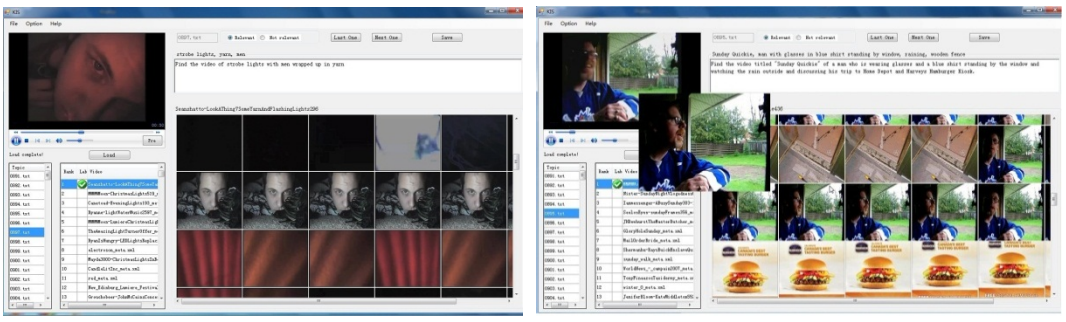
3 Re-ranking

To further improve the performance, we also apply two re-ranking approaches in our system as follows:

(1) **Black/White Video Detection:** Some topics aim at finding a black/white video, such as topic 1181 "*Find the World WarII-era black and white video of the B-60 airplane taking off.*" Thus a black/white video detector [4] is developed to detect whether a video is color or black/white. Firstly, we judge whether a keyframe is black/white. Secondly, if the percentage of black/white keyframes of a video is greater than a threshold, it is regarded to be black/white. Finally, if a topic aims to find a black/white video, all the black/white videos are ranked before the color videos.

(2) **Video Language Detection:** Some topics aim at finding a video in specific language, such as topic 1015 "*Find the video of a seated man with a beard talking in Spanish about 'hipnosis'. Framed documents hang on a green wall behind him. Classical music is playing. A rotating black and white spiral appears.*" Its target is to find a video with "*man with a beard talking in Spanish*". We detect the possible languages of the XML documents by Google Translate. If a topic aims to find a video in a specific language, all the videos with XML documents in this language are ranked before the videos with XML documents not in this language.

4 Interactive KIS System



(a) **(b)**

Figure 3: (a) shows user interface of interactive KIS system, (b) popups an enlarged image when mouse moves onto each keyframe.

An efficient user interface is developed for the interactive KIS system, as shown in Figure 3. A storyboard shows part of the keyframes. The UI popups an enlarged image when mouse moves onto one keyframe, which assists the users to look into the details. Furthermore, if not so sure, the users can view the video and listen to the audio with an embedded windows media player conveniently. After deciding whether a video is correct or not, the users can click "Relevant" or "Not Relevant" to give their opinion.

The UI is simple and intuitive. After reading a topic, the users can reject an irrelevant video just by going through a few keyframes. For example, if the users want to get a video of an airplane, after taking a glance at the keyframes which are all about soccer, they know explicitly it is irrelevant to the topic. In fact, most of the videos are completely irrelevant to a given topic. In this way, a novice can reject an irrelevant video in a few seconds. In our interactive KIS system, 19 of the 24 topics are found, yielding a mean inverted rank of 0.792.

5 Conclusion

By participating in the KIS task in TRECVID 2012, we have the following conclusions: (1) the fusion among varieties of modals information is vital, (2) POS-based keywords extraction, topics term weighting, inflectional normalization and query expansion are key factors, (3) the re-ranking approaches can further improve the search performance.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 61073084, Beijing Natural Science Foundation of China under Grant 4122035, National Hi-Tech Research and Development Program (863 Program) of China under Grant 2012AA012503, National Development and Reform Commission High-tech Program of China under Grant [2010]3044, and National Key Technology Research and Development Program of China under Grant 2012BAH07B01.

References

[1] Lekha Chaisorn, Kong-Wah Wan, Yan-Tao Zheng, Yongwei Zhu, Tian-Shiang Kok, Hui-Li Tan, Zixiang Fu, and Susanna Bolling, "TRECVID 2010 Known-item Search (KIS) Task by I2R", *Proceedings of the 8th TRECVID Workshop*, 2010.

[2] Dan Klein and Christopher D. Manning, "Accurate Unlexicalized Parsing", *Proceedings of the 41st Meeting of the Association for Computational Linguistics(ACL)*, pp. 423-430, 2003.

[3] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System", *Speech Communication*, vol. 37, pp. 89-108, 2002.

[4] Xin Guo, Yuanbo Chen, Wei Liu, Yuanhui Mao, Han Zhang, Kang Zhou, Lingxi Wang, Yan Hua, Zhicheng Zhao, Yanyun Zhao, and Anni Cai, "BUPT-MCPRL at TRECVID 2010", *Proceedings of the 9th TRECVID Workshop*, 2011.

[5] "GNU Aspell", http://aspell.net/

[6] "Kevin's Word List Page", http://wordlist.sourceforge.net/